知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

图书馆海量学术资源自动分类模型研究

◎杨亚 易远弘
宁波大学图书馆与信息中心 宁波 315211

摘要：[目的/意义]针对用户在图书馆海量数字资源中常常面临获取信息困难的问题，构建一套个性化知识服务系统，认为该系统是图书馆帮助用户摆脱信息超载困境和提升知识服务质量的必然选择。[方法/过程]通过建立中图法和学科分类法两大知识组织体系的映射模型，基于 Hadoop 分布式处理平台，提出一种改进 TF-IDF+ 贝叶斯算法构建图书馆海量学术资源自动分类模型，辅助完善图书馆个性化知识服务系统的构建。[结果/结论]以自中国知网抓取的 600 万余篇文献作为原始训练语料(语料涵盖 75 个学科)测试该分类模型的有效性，实验结果证明该模型的分类效率和效果都达到了预期。

关键词：自动分类 Hadoop TF-IDF 算法 贝叶斯
分类号：G250

引用格式：杨亚，易远弘. 图书馆海量学术资源自动分类模型研究 [J/OL]. 知识管理论坛, 2018, 3(3): 172-179[引用日期]. http://www.kmf.ac.cn/p/137/.

随着网络数据库资源和图书馆馆藏数字资源种类和内容的日益丰富，用户经常会在浩如烟海的数字资源中面临获取信息困难的问题。图书馆作为数字资源的再加工者和再组织者，如何有效地组织和管理这些资源，并快速、准确、全面地从中定位到用户所需要的信息是当前图书馆人和信息技术领域面临的一大挑战。自动文本分类是一种处理和组织海量文本资源的有效手段，可在较大程度上解决图书馆文本资源杂乱问题，对于文本资源的高效管理和有效利用都具有极其重要的意义[1]。

基于机器学习的文本自动分类技术，在分类效果和灵活性上都比传统的文本分类模式有所突破，常见的有贝叶斯算法(NB)、k-邻近算法(k-NN)、决策树 (DT)、支持向量机 (SVM) 以及递推神经网络 (RNN) 等[2]。其中贝叶斯分类算法是最常见也是最具代表性的，它是一个基于有监督的机器学习模型，由于其高准确率和高效率一直得到学者们的青睐[3-4]。早在 1998 年 D. Lewis[5]就阐述了如何将贝叶斯应用在信息检索和文本分类领域。后来 Y. LI 等[6]提出一种基于词 - 类别依赖值的加权 NB 算法。

基金项目：本文系浙江省教育厅(文)科研计划项目“面向个性化知识服务的图书馆知识发现平台关键技术研究”(项目编号：Y201635729)和宁波大学科研基金(文)项目“高校图书馆对学生学业成效贡献的大数据分析研究”(项目编号：XYW16004)研究成果之一。
作者简介：杨亚 (ORCID：0000-0002-5758-9464)，工程师，硕士，E-mail: yangya@nbu.edu.cn；易远弘(ORCID: 0000-0002-4616-6052)，助理馆员，硕士研究生。
收稿日期：2018-03-13 发表日期：2018-06-26 本文责任编辑：杜杏叶

知识管理论坛，2018(3):172–179
DOI：10.13266/j.issn.2095–5472.2018.017

邱鹏等[7]提出了一种“先抑后扬”（抑制先验概率的作用，扩大后验概率的影响）的改进贝叶斯文本分类算法。杜选[8]利用类别补集特征消除样本数据分布不均匀，提出一种加权补集的贝叶斯算法。张杰等[9]基于分布式计算框架MapReduce平台，提出一种归一化词频的贝叶斯分类模型。上述学者们在分类算法上做了大量的研究工作，在原始训练语料和实际应用方面相对薄弱，而这正是本文的研究重点。本文基于Hadoop分布式处理平台，通过构建中图法与学科分类法两大知识组织体系的映射模型，采用改进的TF-IDF算法提取文本特征词集，以海量的文本特征词集作为学习语料加入贝叶斯多项式模型进行概率参数训练，完成海量文本数据的并行处理及自动分类模型的构建。

1 图书馆知识组织体系分析

面对海量的学术资源，如何进行有序的组织给我们提出了挑战。本文对原始训练语料的组织策略进行了重点研究，分析整理收集和建立相关知识组织工具，包括主题词表、中图法分类表、学科分类表等，研究学科分类法与中图法两种知识组织体系的内容及关联关系。

学科分类法与中图法是目前对图书馆学术资源进行标引的两大知识组织体系，它们分别从不同内容角度对同一主体进行组织和揭示。中图法是我国图书馆和情报单位普遍使用的一部综合性分类法，主要是供图书馆对图书进行分类管理。中图法包括“马列主义、毛泽东思想，哲学，社会科学，自然科学，综合性图书”5大部类，下一级细分为22个基本大类，每个基本大类又细分为若干门类[10]。教育部颁布的学科分类法一共分为12大学科门类，下一级细分为89个一级学科。中图法的分类太过精细专业化，而学科分类法比较符合用户通常查找资源的习惯。因此，本文选择采用学科分类法对海量的学术资源进行再组织，为更加精准的个性化知识服务提供可靠的保障。

然而，大部分的学术资源没有明确的学科标签，但都带有准确的中图号，因此本文通过手工标引将89个一级学科整理为75个（例如将“理论经济学”与“应用经济学”合并称为“经济学”，“地理学”与“大气科学”合并称为“地球科学”），然后建立75个一级学科与中图法的22个基本大类的一一映射表（见表1），构建两大知识组织体系的关联模型，为后期训练分类模型做好基础准备。

表1 中图法与学科分类法部分映射表

中图号	学科名称	中图号	学科名称
A1~A8	马克思主义理论	C94	系统科学
B0~B83,B9	哲学	C95	民族学
B84	心理学	D0-D8	政治学
C0-C7,C91-C92	社会学	D9-DF	法学
C8	经济学	E0-E9	军事学
C93,C96-C97	管理科学与工程	…	…

2 图书馆海量学术资源自动分类模型构建

图书馆学术资源自动分类模型的分类效果很大程度上依赖于原始训练语料的质量和总量。本文的原始训练语料主要是抓取自中国知网的600万篇以上的高质量语料，语料涵盖75个学科。利用学科分类法与中图法关联模型，基于Hadoop分布式处理平台，对图书馆海量学术资源进行批量的训练与分类。包括三步（见图1）：

第一步：数据的预处理。提取分类所需的关键字段，包括题目、摘要、关键词、中图号等。然后对语料库进行分词、去停留词、保留专有名词。

第二步：提取文本关键词集。引入TF-IDF算法对原始词集进行关键词提取，并作为学习语料放入分类模型进行训练。

第三步：贝叶斯多项式分类模型训练。将文本关键词集作为输入特征分别计算该篇文本可能归属学科类别的概率值，选取概率值最大的类别作为该篇文本的类别。

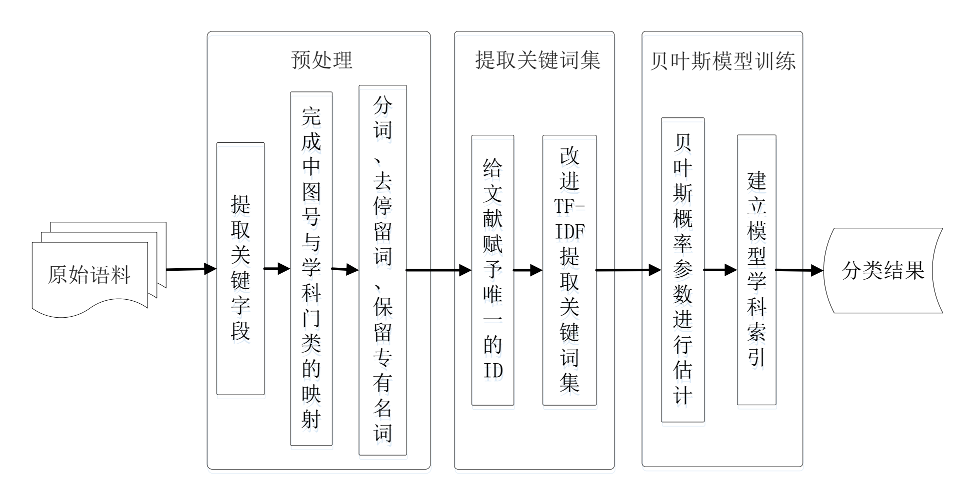


图 1 海量学术文献自动分类流程

2.1 数据预处理

在数据预处理阶段主要是完成对原始语料关键字段的提取及分词操作。题目、摘要、关键词和中图号是一篇学术文献最能体现其主题与所属学科的关键要素，因此，本文提取这几个关键字段构建原始语料库。其中，提取中图号是与第 1 小节的关联模型进行学科类别匹配。

接下来是对原始语料库进行分词过滤操作。目前开源的分词器很多，IK 分词器是一个基于 Java 语言设计开发的中文分词工具，它自身带有停用词表，可在其中添加自定义的停用词[11]。由于本文处理的数据都是学科类数据，数据中包含比较多的专有名词，IK 分词器正好满足这样的需求，可以自行添加专有名词表，避免专有名词被切分。本实验添加了 20 万条学术专有名词词库，基于 MapReduce 框架实现并行化分词的步骤如下：

第一步：自定义输入类 SubjectInputFormat 和一个 paths 数组，SubjectInputFormat 继承 FileInputFormat 类并重载 getSpilts 方法实现多个文件分片，paths 数组用于记录每个文件的路径；

第二步：定义一个构造函数 SubjectRecordReader 来处理分片内容，通过 SubjectInputFormat 调用 CreateRecordReader 方法并返回 CombineRecordReader 对象，将结果 <Key,Value> 对传递到 Mapper 中。其中 Key 代表文件所属的类别名，Value 代表文件内容，类型均为 Text.。

第三步：Mapper 端接收到 Value 中的文件内容后调用IK 分词器提供的接口进行分词处理。

2.2 提取文本关键词集

为了提高文本分类的效率和准确度，本文引入经典的 TF-IDF 算法对原始词集进行关键词的提取[12,13]。TF 表示某一词汇在文本中出现的频率，IDF 表示逆文本频率（能够反映该词在整个语料库中的大众化程度），文本中每个词都可以通过这两个指标的乘积得到一个权重即 *tf*idf* 值，按一定的比例筛选出权重较大的词作为该篇文本的关键词集。如下列公式所示：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \text{公式（1）}$$

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \qquad \text{公式（2）}$$

公式（1）中，$n_{i,j}$ 表示某一词汇 t_i 在该篇文本中出现的次数，分母为该篇文本中所有词汇出现次数的总和。公式（2）中 $|D|$ 表示语料库的文本总数，$|\{j:t_i\in d_j\}|$ 表示包含词汇 t_i 的文本数目，如果该词汇不在语料库中，就会导致被分母为零，所以一般情况使用

知识管理论坛，2018(3):172–179
DOI：10.13266/j.issn.2095–5472.2018.017

$\left|\left\{j : t_i \in d_j\right\}\right| + 1$。

根据算法中 IDF 的定义，当词汇 t_i 在某个学科 b_l 频繁出现，而在其他学科极少出现时，通常会被赋予较低的权重，说明该词汇类别区分能力不强。但是实际上，当学科 b_l 中包含词汇 t_i 的文本数量大，而其他学科中包含词汇 t_i 的文本数量小，则说明词汇 t_i 能很好地代表学科 b_l 的文本特征，具有很好的类别区分能力。因此，针对这个缺陷，张玉芳[14] 等将公式（2）变形为

$$idf_i = \log\left(\frac{\left|\{l : t_i \in b_l\}\right|}{\left|\{l : t_i \in b_l\}\right| + \left|\{m : t_i \in c_m\}\right| + 1} \times |D|\right)$$

公式（3）

$\left|\{l : t_i \in b_l\}\right|$为修正系数，表示某个学科 b_l 中包含词汇 t_i 的文本数目，$\left|\{m : t_i \in c_m\}\right|$表示除学科 b_l 外包含词汇 t_i 的文本数目。可以看出 idf_i 随着$\left|\{l : t_i \in b_l\}\right|$的增大而增大，随着$\left|\{m : t_i \in c_m\}\right|$的增大而减小，这刚好能弥补 IDF 定义的缺陷。

张玉芳等提出的方法对权重修正有一定的作用，在此基础上，本文还考虑特征词在文本中不同位置的类别区分能力不同，引入一个位置因子 σ，特征词出现在关键词位置应具备更好的类别区分能力，大量实验显示 σ=2 时效果比较好，则公式（3）变形为

$$idf_i = \sigma(t_i) \times \log\left(\frac{\left|\{l : t_i \in b_l\}\right|}{\left|\{l : t_i \in b_l\}\right| + \left|\{m : t_i \in c_m\}\right| + 1} \times |D|\right)$$

公式（4）

此外，在一篇文献中，每个词的 TF（频度）的计算量与文本长度成正比，而 IDF（逆文本频率）的计算量则与语料库的大小成正比。本文初始语料库中包含语料 600 万篇以上，而且会语料库会不断地更新，如果每次计算 *idf* 值时都去基于语料库统计，响应时间较长且浪费计算资源。因此，本文基于 MapReduce 事先对整个语料库中的每个词计算出 *idf* 值，并将结果存放在 Mysql 数据库中，当计算某个词汇的 *tf*idf* 值时，直接去 Mysql 数据库中取即可。实现过程如下：

步骤一：给语料库中每篇文献赋予一个唯一的标识 id。

步骤二：定义 mapper<key,value> 函数，key 为分区字节偏移量，value 为 < 文本 id>|< 分词后的词集 >|< 学科类别 >。对词集进行迭代输出，输出 key 为 < 特征词 >|< 学科类别 > 形式的字符串，value 为文本 id。

步骤三：定义 reducer<key,value> 函数，key 为某一词汇 < 特征词 >|< 位置因子 >|< 学科类别 > 形式的字符串，value 为该词汇对应的所有文本 id。首先对同一词汇下的文本 id 进行去重并计算出包含该词汇的文本数量 n，然后统计出所有类别中含该词汇的文本数目，文本数目最大的记作 m，最后计算该词的 idf 值即 $idf = \sigma \times \log\left(\frac{m}{n+1} \times |D|\right)$。输出 key 为某一词汇，value 为该词汇对应的 idf 值。将结果放入 Mysql 数据库中，并建好索引。

步骤四：提取特征词集，首先计算某个词汇在文本中出现的频率 tf，然后从数据库中读取对应词汇的 idf 值，即可得到该词的 *tf*idf*，按照一定比例选取该文本的特征词集。

2.3 贝叶斯多项式分类模型训练

贝叶斯文本分类算法的理论基础是假设组成文本的词汇之间是相互独立的，在先验概率和条件概率的基础上计算最终的后验概率，选取概率最大作为分类的结果。

给定文本训练集（x,y），某个文本有 n 个特征词，即 $x=(x_1,x_2,\ldots,x_n)$，每个特征词有 k 种类别，即 $y=(y_1,y_2,\ldots,y_k)$，则分类函数记作 $f(x) = \arg\max_{y_k} P(y_k \mid x)$，即转化为求解概率函数 $P(y_k|x)$，即

$$P(y_k \mid x) = \frac{p(x \mid y_k)p(y_k)}{p(x)} = \frac{p(x \mid y_k)p(y_k)}{\sum_k p(x \mid y_k)p(y_k)}$$

公式（5）

在公式（5）中，对于所有的 y_k，分母的值都一样，所以可以忽略分母部分；$P(y_k)$ 是先验概率，根据训练集就可以简单地计算出来；然后根据贝叶斯理论假设特征词属性 $x_1,x_2,\ldots,x_n$ 互相独立，则



$$P(x|y_k)=P(x_1,x_2,\cdots,x_n|y_k)=\prod_{i=1}^{n}P(x_i|y_k)$$ 公式（6）

那么最终贝叶斯分类函数表示为：

$$f(x)=\arg\max_{y_k}P(y_k)\prod_{i=1}^{n}P(x_i|y_k)$$ 公式（7）

由于待分类文本的特征是离散的，使用多项式模型来计算先验概率和条件概率，公式如下：

$$P(y_k)=\frac{N_{y_k}+\alpha}{N+k\alpha}$$ 公式（8）

$$P(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+M\alpha}$$ 公式（9）

N_{y_k}是类别为 y_k 的文本数量，N 是总的样本数量，N_{y_k,x_i}是类别为 y_k 的文本中，包含特征词 x_i 的文本数量，M 表示训练集中所有关键词的数量，α 是平滑值，在实际应用中一般取值 1。

在贝叶斯分类模型训练阶段，主要任务就是对参数 $P(y_k)$ 和 $P(x_i|y_k)$ 进行估计，前者是对训练集的类别统计，后者则需要基于语料库统计每个词与 75 个学科类别的关系。因此，使用 MapReduce 编程框架实现基于海量语料库的贝叶斯模型参数估计，并将各个参数值放入 Mysql 数据库，以便模型进行学科标引时使用。具体步骤如下：

第一步：定义 mapper<key,value> 函数，key 为分片偏移量，value 为 <文本 id>|<关键词集>|<对应的类别标签>。然后针对每个关键词进行输出，输出 key 为 <关键词>|<类别>形式的字符串，value 为文本 id。

第二步：定义 reducer<key,value> 函数，初始化时加载各类别文章在语料库中的数量和语料库中不重复的关键词数 M。输入的 key 为 <关键词>|<类别>形式的字符串，value 为文本 id 的组合。该函数主要是对 <关键词>|<类别>对应下的文本 id 集合进行去重并计算总数 n，并获取该类别的文本数 m，输出的 key 为 <关键词>|<类别>形式的字符串，value 为 (n+1)/(m+M) 的比值。

第三步：将处理结果导入到 Mysql 中，并建好索引，要对某一文本进行分类时，直接去数据库中取相应的值进行计算即可。

3 实验与分析

3.1 分类评价指标

文本分类器常采用的评价指标是查准率 P、查全率 R 以及基于两者的综合指标 $F1$，查准率 P 是分类器正确判断为该类的样本数与判断属于该类的样本总数的比率，查全率 R 是分类器正确判断为该类的样本数与属于该类的样本总数的比率 [15]。计算公式如下：

$$P=\frac{a}{a+b}$$ 公式（10）

$$R=\frac{a}{a+c}$$ 公式（11）

a 为属于某类别且被判定为该类别的文本数量，b 为不属于某类别但被判定为该类别的文本数量，c 为属于某类别但未被判定为该类别的文本数量。根据 P 和 R 计算出 $F1$，即

$$F1=\frac{P\times R\times 2}{P+R}$$ 公式（12）

3.2 实验结果与分析

实验训练及测试使用的语料均抓取自中国知网近三年的论文数据，语料涵盖了所有的学科类别。为了验证本文分类模型的有效性，本文进行了 2 个实验：

（1）传统 TF-IDF 算法与改进 TF-IDF 算法计算 idf 值的对比实验。随机选取 200 个特征词的 idf 值对比情况，部分结果见表 2。

通过大量的实验发现，经过改进 TF-IDF 算法计算得到的 idf 值都有一定的变化，“临床”“患者”等的 idf 值变化比较大，具有较强的类别区分能力，由此说明改进的 TF-IDF 算法能提高某些学科专有名词的权重。

（2）LDA+SVM 与改进 TF-IDF+ 贝叶斯两种分类策略的对比实验。为消除样本不平衡对分类结果的影响，在实验之前随机选取 20 个类别，每个类别随机抽取 2 000 条数据作为训练集，1 000 条数据作为测试集，训练集与测试集的数量比为 2:1，且没有重复数据。实验部分结果见表 3。

知识管理论坛，2018(3):172–179
DOI：10.13266/j.issn.2095–5472.2018.017

从实验结果可以看出，相比 LDA+SVM 分类算法，采用改进 TF-IDF+ 贝叶斯算法得到的查准率、查全率以及 F1 值都有明显的提高。此外，本文基于 Hadoop 分布式处理平台构建的自动分类模型，相较于在单机上实现，计算效率得到了大大的提高。因此，可以得出本文构建的自动分类模型在处理海量文本分类时具有一定的优势。

表 2 idf 值对比情况

特征词	TF-IDF 算法	改进 TF-IDF 算法	特征词	TF-IDF 算法	改进 TF-IDF 算法
仮佛	5.804	5.804	观察	1.207	1.906
新牌	5.615	5.615	利用	1.150	1.817
前庭大腺	4.976	5.277	相关	1.116	1.327
复兴路	4.858	5.050	作用	1.097	1.385
镇肝熄风汤	4.854	5.133	临床	1.092	2.372
菠萝蜜	4.849	5.060	中国	1.069	1.294
腊肠	4.620	4.958	进行了	1.043	1.224
剖面图	4.364	4.630	患者	0.993	2.178
补中益气	4.219	4.616	影响	0.879	1.191
诺尔	4.208	4.519	结论	0.851	1.136
血液制品	4.206	4.600	目的	0.809	1.054
上海大学	4.086	4.480	本文	0.787	0.984
味觉	3.978	4.304	发展	0.773	0.918
重点企业	3.834	4.292	研究	0.758	0.909
...	...	...	分析	0.569	0.879

表 3 实验结果

学科类别	LDA+SVM			改进 TF-IDF+ 贝叶斯		
	P	R	F1	P	R	F1
天文学	0.798	0.756	0.776	0.855	0.847	0.851
控制科学与工程	0.752	0.748	0.750	0.828	0.857	0.842
经济学	0.749	0.729	0.739	0.821	0.846	0.833
计算机科学与技术	0.798	0.762	0.780	0.839	0.858	0.848
数学	0.720	0.736	0.728	0.834	0.845	0.839
军事学	0.756	0.764	0.760	0.825	0.846	0.835
水产	0.781	0.758	0.769	0.836	0.845	0.840
食品科学与工程	0.769	0.786	0.777	0.849	0.861	0.855
...	...	...	...	...	...	...

4 结束语

图书馆每年都有大量的学术资源产生，在面对这些资源时，用户总是陷入获取精准资源的困境。如何帮助用户找到既优质又相关的学术资源是图书馆亟待解决的一项实际问题。本文对基于机器学习的文本自动分类技术进行研究，旨在通过对海量学术资源进行高效而准确

的自动分类，辅助构建图书馆个性化知识服务系统。

本文在 Hadoop 分布式处理平台上，对图书馆海量的学术资源进行并行化处理，大大的提高了计算的效率。提出一种改进的 TF-IDF 算法进行文本特征词集提取，既能过滤掉大量的噪音词汇，降低计算的复杂度，也便于后面对海量的学术资源进行贝叶斯分类处理，提升分类的准确度。存在的不足之处是分类模型采用的是批量处理模式，即批量的训练和批量的分类，尚未做到增量训练和实时分类。接下来的工作可考虑基于 Spark 搭建流处理平台，以及研究如何实现模型的增量学习。

参考文献：

[1] VIKAS K, VIJAYAN K, LATHA P. A comprehensive study of text classification algorithms[C]// Proceedings of 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Udupi:IEEE press, 2017: 1109-1113.

[2] 高元 . 面向个性化推荐的海量学术资源分类研究 [D]. 宁波 : 宁波大学 , 2017.

[3] 贺鸣 , 孙建军 , 成颖 . 基于朴素贝叶斯的文本分类研究综述 [J]. 情报科学 , 2016, 34(7): 147-154.

[4] KUPERVASSER O. The mysterious optimality of naive bayes: estimation of the probability in the system of “classifiers”[J]. Pattern recognition and image analysis, 2014, 24(1): 1-10.

[5] LEWIS D. Naive (Bayes) at forty: The independence assumption in information retrieval[C]//Proceedings of 10th European Conference on Machine Learning Chemnitz. Berlin: Springer, 1998: 4-15

[6] LI Y J, LUO C N, CHUNG S M. Weighted naive bayes for text classification using positive term-class dependency[J]. International journal on artificial intelligence tools, 2012, 21(1): 1250008-1250015.

[7] 邸鹏 , 段利国 . 一种新型朴素贝叶斯文本分类算法 [J]. 数据采集与处理 , 2014, 29(1): 71-75.

[8] 杜选 . 基于加权补集的朴素贝叶斯文本分类算法研究 [J]. 计算机应用与软件 , 2014, 31(9): 253-255.

[9] 张杰 , 陈怀新 . 基于归一化词频贝叶斯模型的文本分类方法 [J]. 计算机工程与设计 , 2016, 37(3): 799-802.

[10] 艾雰 . 2010—2016 年《中国图书馆分类法》(第 5 版) 研究现状分析 [J]. 图书馆建设 , 2017(5): 39-44,72.

[11] LI Q, CHEN L. Study on multi-class text classification based on improved SVM[C]//Proceedings of the Eighth International Conference on Intelligent Systems and Knowledge Engineering, Shenzhen: Springer Berlin Heidelberg, 2014: 519-526.

[12] ZHANG Y T, WANG G L. An improved TF-IDF approach for text classification[J]. Journal of zhejiang university-science a, 2005, 6(1): 49-55.

[13] KIM S B, RIM H C. Effective Methods for improving naive bayes text classifiers[C] //Proceedings of 7th Pacific Rim international conference on artificial intelligence. Berlin:Springer, 2002: 414-423.

[14] 张玉芳 , 彭时名 , 吕佳 . 基于文本分类 TFIDF 方法的改进与应用 [J]. 计算机工程 , 2006(19): 76-78.

[15] 苏金树 , 张博锋 , 徐昕 . 基于机器学习的文本分类技术研究进展 [J]. 软件学报 , 2006(9): 1848-1859.

作者贡献说明：

杨　亚：数据处理，撰写并修正论文；

易远弘：学术资源整理，提出论文的修改意见。

知识管理论坛，2018(3):172–179
DOI：10.13266/j.issn.2095–5472.2018.017

Research on Automatic Classification Model of Massive Academic Resources in Library

Yang Ya Yi Yuanhong
Library and Information Center, Ningbo University, Ningbo 315211

Abstract: [Purpose/significance] In order to solve the problem that users often have difficulty in obtaining information in massive digital resources of library, this paper construct a personalized knowledge service system, which is the inevitable choice of library to help users to get rid of the information overload predicament and improve the quality of knowledge service. **[Method/process]** Firstly, this paper built a mapping model of Chinese Library Classification(CLC) and subject classification. Then, based on Hadoop distributed processing platform, it proposed to build automatic classification model of massive academic resources in libraries by improving TF-IDF+ Bayesian algorithm, the model can help to construct the personalized knowledge service systems in library. **[Result/conclusion]** In the experimental part，we collected more than 6 million documents from CNKI as the original training corpus (corpus covers 75 disciplines) to test the effectiveness of the classification model, the experimental result shows that the classification efficiency and effectiveness of the model are achieved.

Keywords: automatic classification Hadoop TF-IDF Bayes